CHAPTER 2: BACKGROUND

Prevalence and Nature of High School Graduation Testing

Testing is being used by schools and states for a wide range of purposes from diagnosing student deficiencies to holding schools accountable for student achievement. The focus of this paper is on high school exit exams.

High school exit examinations present considerable issues for state policymakers because of their "high stakes" nature. Important consequences are attached to test results, and an exit exam is high stakes in nature because award of a diploma is based on passing the test (NGA Reports Online, 1998). In fact, *High Stakes* is the leading title used by the National Research Council's Committee on Appropriate Test Use to report on the Congressionally-mandated study on uses of testing for tracking, promotion, and graduation (National Research Council, 1999). The courts have held that a high school diploma is a property interest, and as such, protected by the Fourteenth Amendment. As a result, implementation of an exit exam is likely to present state policymakers and assessment staff with technical and legal challenges (National Research Council, 1999; NGA Reports Online, 1998).

According to the Council of Chief State School Officers' (CCSSO) annual survey of state student assessment programs, eighteen states had implemented high school exit examinations as of the 1997–98 school year (Olson, Bond & Andrews, 1999). Table 2.1 lists the states and indicates their responses to a survey question asking to what instructional uses these assessment results were put.

The number of states with such a requirement is on the rise; NGA Reports Online (1998) identifies twenty. These range from Florida, which first implemented its test in the 1977–78 school year (then delayed until the 90s as significant legal challenges occurred), to Alaska, whose exam requirement becomes effective in 2000 (NGA Reports Online, 1998). In most of these states, students must pass the examination in order to graduate from high school; other states offer an honors or endorsed diploma based upon test results (Olson et al., 1999; NGA Reports Online, 1998). Passing the test is typically linked with completing proscribed coursework; that is, a student must complete all required coursework and pass the exit examination in order to receive a high school diploma.

Half of the participating states test students on material that is typically taught in 9th grade or earlier; these tests can be regarded as minimum competency tests (National Research Council, 1999). Other states have implemented tests containing more advanced material. There is, however, a degree of tension between minimum competency or basic skills approaches and the more recent push for all students to learn at higher levels. Many states initiated state standards after implementing a graduation test, so policymakers need to know that there may be problems of alignment between the content of the exam and the content of the standards (NGA Reports Online, 1998). Even with this caution in mind, Boser (2000) reported preliminary results from a University of Wisconsin study of 10 states that claims there may be little overlap between what state assessments test and what teachers teach—with a range from 5% to 46%. This adds to the existing concerns about standards-based accountability; although teachers are likely to be teaching what they think is appropriate,

their individual lesson plans may not be aligned with state standards or assessments (Boser, 2000).

Table 2.1. States with high school exit examinations as of 1997–98.

		Instructional Uses					
State [1]	Implementation Dates	Identification of students at risk	Student diagnosis	Individual student instructional planning	Improvement of instruction for groups of students	Curriculum planning at school/ district	Program evaluation
Alabama	March 1999: Class of 2001		✓	✓			
Florida	October 1993: Class of 1995	✓		✓	✓	✓	
Georgia	Fall 1993; Class of 1995	✓	✓	✓	✓	✓	
Hawaii	Spring 1979; Class of 1983	✓	✓	✓	✓	✓	
Indiana	September 1997; Class of 2000	✓	✓	✓	✓	✓	✓
Louisiana	March 2001 & 2002; Class of 2003		✓	✓		✓	
Maryland	1978–79; Class of 1982[2]	✓	✓	✓			
Mississippi	Fall 1987; Class of 1989						
Nevada	April 1998; Class of 1999	✓	✓	✓		✓	
New Jersey	October 1992; Class of 1994	✓	✓	✓	✓	✓	✓
New Mexico	February 1988; Class of 1990						
New York	1987–88[3]						
North Carolina	Summer 1994; Class of 1998	✓			✓	✓	✓
Ohio	Fall 1990; Class of 1994				✓	✓	
South Carolina	Spring 1986; Class of 1990[4]		✓		✓		✓
Tennessee	Phasing out	✓	✓	✓	✓		
Texas	2002–2003	✓	✓	✓	✓	✓	✓
Virginia	Spring 1998; Class of 2004	✓	✓	✓	✓	✓	✓

Policymakers should anticipate that the approach that calls for more rigorous assessments is likely to produce higher failure rates, particularly among educationally disadvantaged

[1] The District of Columbia plans to implement an exit exam starting with the Class of 2003; and Massachusetts' plan starts with the Class of 2003, which must pass 10th grade assessment (Argetsinger, 2000).
[2] Maryland functional Testing Program is currently being phased out, to be replaced by a new high school assessment program in January 2002, Class of 2005. However, Maryland recently voted to delay its high school exit exam by at least two years; the decision means the Class of 2007 is likely to be the first to face the test as a graduation requirement (Argetsinger, 2000).
[3] New York's Regents Competency Tests are currently being phased out in favor of Regents Examinations as of June 1999, Class of 2005.
[4] South Carolina's Basic Skills Assessment Program (BASP) is being replaced by Palmetto Achievement Challenge Tests (PACT), Spring 2001, Class of 2003.

students. High failure rates, in turn, may be expected to increase public concern and the likelihood of legal challenges. The Maryland decision to slow the development and implementation timeline of its assessment program points out the public and political concerns that can arise over the consequences of testing rigor (NGA Reports Online, 1998). The Maryland State Board of Education has voted to delay the graduation test requirement by at least 2 years because of added concerns that the state needs to do more to help children prepare for the challenge (Argetsinger, 2000; Solida, 2000). This parallels a caution in the National Academy of Education report (McLaughlin, Shepard, & O'Day, 1995) that before all students are expected to perform to world-class standards, they must be provided a world-class curriculum and instruction. This is likely to be a particular concern in California, where the high standards reflected on the graduation test currently are a challenge to the state's lack of authority to tell districts what courses to teach, and the low per-pupil expenditures are a challenge to upgrading and aligning curriculum and instruction to a level sufficient to meet the expectation the test creates.

For minority children, English-language learners, and students with disabilities, the stakes in appropriate, nondiscriminatory test use are particularly high. Such children tend to do less well on large-scale tests, even where such tests measure only basic skills. In the late 1970s, for example, when minimum competency tests gained popularity, 20% of African American students, compared with 2% of White students, failed Florida's graduation tests and were denied high school diplomas (*Debra P. v. Turlington*, 1979). Similarly, 1998 data from the Texas graduation tests show cumulative failure rates of 17.6% for African American students and 17.4% for Hispanic students, compared with 6.7% for White students (Natriello & Pallas, 1999). For students with disabilities, most of whom now participate in large-scale state assessments, recent data from 14 states consistently show failure rates that are 35–40 percentage points higher than those for nondisabled students (Ysseldyke et al., 1998).

Moreover, states increasingly are raising the standards on their graduation tests to "world-class" levels, such as those embodied in the National Assessment of Educational Progress (NAEP). Based on 1996 NAEP data, 40% of all students would fail tests that reflect world-class" standards, at least initially, and failure rates for minority students would be about 80%. These predictions are consistent with recent data from Massachusetts and New York, where students have begun taking state tests that reflect world-class standards. On such tests, failure rates for students with disabilities would likely be in the range of 75 to 80%.

Regardless of whether the exam content addresses basic skills or more challenging "world-class" levels, statewide high school exit exams are intended to provide a consistent, objective measure of student learning throughout the state. The most commonly tested subject areas are language arts (English reading and/or writing) and mathematics; some states also test science and social studies (NGA Reports Online, 1998). The most common item format in large-scale assessment programs is multiple choice. Most states use multiple choice in conjunction with other assessment types such as extended constructed response, short answer, and performance items (Olson et al, 1999). All of the states with exit exams allow students to retake the test multiple times.

The Individuals with Disabilities Act (IDEA) and the Improving America's Schools Act (IASA, 1994) demand that all students be included in state assessments and that all students be offered equal educational opportunities and be held to the same high standards. According to the IASA, Title I programs must have in place final assessment systems that include all students and provide disaggregated (by component categories such as gender and race/ethnicity) test scores by school year 2000–2001.

Policymakers clearly are going to have to address such requirements and concerns, as well as deal with a number of other policy issues. One concern is that graduation tests may run counter to motivating students, especially high achieving ones. The fear is that students will view the "minimum" level of knowledge and skills as the "maximum" they need to master in high school. Another issue is timing of the test. When administered initially in 8th or 9th grade, the content cannot include material taught at later grades. Thus the test cannot be as rigorous as if it were to be administered later. However, when tests are administered later, students who do not pass have limited time for remediation. Some states have established "early warning tests" to serve as alerts of any student learning deficiencies. Remediation itself is probably one of the greatest challenges for states. Policymakers, who have a responsibility in overseeing all aspects of the testing program, need to make sure that all appropriate steps are taken to help students who fail pass the test (NGA Reports Online, 1998).

From Testing to Higher Achievement

For proponents of such testing, high school exit exams have an immediate goal: to raise the standard by which a high school diploma is awarded. A long-term implication of this higher standard would be greater achievement for students currently below this level. Consistent with this implicit goal, the 1999 Council of Chief State School Officers' (CCSSO; Olson et al., 1999) survey reports that the most common reason cited by states for their various assessments, not limited to the high school exit exam, was to improve instruction. School accountability, student accountability, and staff accountability were cited with far less frequency.

There are two routes by which a high school exit exam may impact student achievement. First, student motivation may be affected. Second, instruction can be modified, presumably for the better.

The effect of a high stakes exam on student motivation is a double-edged sword. The beneficial aspect is that student motivation may increase when students and parents realize that the exam must be passed in order to earn a high school diploma. High achievers may increase effort in order to pass the exam as soon as possible, while students who fail the exam on the first attempt may see this as a wake-up call and try harder. On the other hand, some argue that the exam may be a demotivator for students. They warn that high-achieving students who are able to pass the exam early may lose motivation to continue to strive for further learning. Since many states tie their exam contents to 9th grade curricula (e.g., Algebra), this is a viable concern. Further, they claim that students who fail the exam may become disillusioned with education and their abilities (CCSSO; Olson et al., 1999).

The link between mandatory testing and improved instruction, and by implication, higher achievement, is not a direct one. By themselves, tests improve education no more than a thermometer reduces fever—the existence of a test ensures nothing other than that the level of achievement will be measured. In order for a high stakes statewide testing program to foster a higher level of student achievement, several intervening connections must be in place. Like the links in a chain, each connection is an integral part of a fully-functioning system.

For example, if test results are to accurately reflect achievement, the test must meet standards of appropriate test use (e.g., validity, psychometrics, content). In order for test results to subsequently affect achievement, test scores, or the framework used in building the test, must provide information that guides instructional changes; therefore, the test should lead to the identification of specific areas of instruction that need shoring up. Further, in order for this diagnostic information to be put to use, the score reports must be interpretable. Even more fundamentally, the relevant parties must be aware of the information available in the score reports. Not only must test results be diagnostic, understandable, and publicized, but educators must be able to determine what the results imply for current practice and enjoin for future practice. At a broader level, the aggregation of test results across schools, districts, and student demographic groups should provide useful information regarding equity of successful instruction. Tools and procedures to monitor and rectify any inequities must be in place before any high stakes student impact is enacted.

The following sections address each of these links between high stakes testing and improved instruction.

Standards of Appropriate Test Use

The National Research Council's committee on high stakes testing relied upon three broad principles of appropriate test use (National Research Council, 1999). First, the test should have measurement validity (i.e., be valid for its stated purpose), defensible test psychometrics, and appropriate content. In order for an examination score to be a mandatory component of a high school diploma, the content must be included in curricula throughout the state. Any inconsistency opens the door for legal action from students claiming not to have been taught the requisite material. Before or after the fact, the presence of a mandatory statewide exam presupposes commonality among all schools throughout the state. If not already present, one of the consequences of such an exam is the alignment of curricula with exam contents. The *Debra P. v. Turlington* case (1981) established, among other things, that graduation test contents must reflect what students have been taught. In reporting on a forthcoming study from the University of Wisconsin, Boser (2000) reiterates the concern on basic premise of implementation: "States should delineate what students should know and be able to do, teachers should match instruction to those standards, and state tests should measure how well students meet those expectations." However, the Wisconsin researchers found some criticism of states for not doing enough to explain to teachers what is on the exam, given the penalties and rewards systems that are increasingly accompanying the tests (Boser, 2000). In California this is likely to be exacerbated by the state's limited authority to tell districts what to teach.

Second, test users must make valid attributions of cause. Poor scores can be the result of students simply failing to learn what they've been taught, or confounded by other factors such as disabilities or language barriers that are not related to the construct, or a denial of opportunity to learn.

Third, the test should address consequences (consequential validity). In other words, the test should result in better education.

Diagnostic Information

The specificity of a test score impacts the utility of that score. For example, a low score in mathematics may be the result of any of a number of specific deficits. If only a single mathematics score is reported, this may not provide sufficient detail to instigate appropriate remedial measures. In the absence of diagnostic scores for individual students, school faculty, parents, and students are left guessing as to what material or knowledge is lacking in a less-than-satisfactory score. Not only might this result in inappropriate or unnecessary remedial instruction for the individual student, but it might also fail to illuminate school- or districtwide curriculum gaps, as well as differential results for particular groups of students (e.g., minority students, low-socioeconomic (SES) students, English-language learners, students with disabilities). Without diagnostic-level information, an unintended consequence may be the ill-informed adoption of ineffective "quick fixes" to raise scores.

This is an issue in California as discussion continues regarding whether the exit examination is intended as a certification exam or as a basis for remediation. If the plan is to allow students four times to pass the test, it would seem that the year between each testing window is intended for students and teachers to work on the problem areas to increase the likelihood of passing in the future. If the time between test administrations is not meant for teaching and learning to promote passing, but rather only for certification, then this process must be carefully and clearly explained.

The specificity of reliable assessment stipulates test length, administration time, and cost, among other things. These factors may preclude the generation of a full level of diagnostic detail by the statewide exit exam. For practical purposes, the exit exam score may be limited to aggregate levels (i.e., pass/fail math). If so, other sources of information (e.g., other test scores, teacher evaluation) may be recruited at the schoolhouse level to augment the score for diagnostic purposes.

Meaningful Score Reports

In order to be put to good use, the score reports must be clear to educators. At both the individual student level and the aggregate classroom/schoolhouse level, teachers must be able to understand the implications of the reports. This requires, at a minimum, a well-designed report; it may additionally call for professional development training and score interpretation guides.

Awareness and Communication

An easily overlooked aspect of a new testing system is the propagation of information, including counsel about what data are available. For example, if a school-level diagnostic

report is provided along with individual student results, educators must be aware of the information provided by the school-level report and how to use it, in order for it to be utilized effectively. If data and tools are available to permit schools/districts to aggregate data at various levels, personnel must be trained and procedures put in place to facilitate their use.

The 1998 CCSSO survey asked of each state, "What kinds of [successful] materials has your state or others in your state developed for assessment program publicity, explanation, or training?" Frequent responses included training manuals, newspaper inserts, brochures, web site information, videos, public access television, and interpretive guides. Most states also reported various forms of professional development to educate teachers and other educational staff on the statewide assessment program. These ranged from paper documents to hands-on workshops (Olson et al, 1999).

Identification of Appropriate Instructional Changes

Raising awareness about and understanding of test results are necessary steps toward eliciting appropriate instructional changes, but they are not enough. School personnel must be able to determine what various patterns of results imply for instruction, such as which instructional units are succeeding and which are falling short. Areas needing modification may include changes such as additional instructional time for certain concepts, inclusion of additional concepts, or altered pedagogical approaches. These analyses should be conducted at the individual level, as well as at various aggregate levels (e.g., English-language learners, students with disabilities, schoolwide, districtwide).

Implementation of Appropriate Instructional Changes

The application of test results to appropriate instructional changes relies upon teacher training and availability of adequate resources in a sufficient timeframe. Teachers must know what adjustments to make, how to make them, and have the necessary time, money, and facilities to implement those changes. This is a two-pronged process: remediation of students who failed the exam already, and modification of first-time instruction to students who have not yet taken the exam.

In 1994, 7 of the 18 states with graduation examinations provided funds to schools or districts specifically for remedial education (Bond & King, 1995). Another approach to remediation is to provide additional educational material directly to students in need. For the Texas Assessment of Academic Skills (TAAS), the state provides *TAAS Study Guides* free of charge to school districts, concurrent with test results. These are distributed to students who fail any parts of the TAAS. Connecticut, Kentucky, and Maryland are other states that have spent heavily to bring teaching and learning into line with state standards.

Modification of first-time instruction presumes feedback to the teachers of subjects being tested. For example, if the students in a particular high school have difficulty with the Algebra portion of the exam, feedback should be provided to the middle schools that feed that high school, as some of the students would have completed Algebra in the 8th grade. Algebra instruction may require modification at both the middle and high school levels. In order for this to happen, there must be a feedback mechanism within the high school, as well as from the high school to its feeder schools.

Modification of instruction to enhance test scores brings with it short- and long-term consequences, some good and some bad (Hess & Brigham, 2000). Long-term benefits include increased equity, clarity, and focus in curriculum, and an efficient use of resources. These benefits come hand in hand with consequences such as limited local decision-making and a narrowed curriculum. Despite the long-term equity advantages, in the short term there are negative consequences for poor, minority, and special needs students. Chudowsky and Behuniak (1998) also stress the negative consequences of a narrowed curriculum, as well as the loss of instructional time to preparation and administration of the test. Their teacher focus groups also revealed classroom benefits such as the availability of test practice materials, new instructional approaches, and the purchase of more lab equipment and textbooks.

Anticipated Challenges to California's High School Exit Exam

The history of public education in the United States is marked by numerous legal challenges. The U.S. Constitution, federal civil rights statutes, and other judicial decisions have been invoked to determine whether specific tests are discriminatory or otherwise inappropriate. In recent years, high stakes testing programs have faced complaints of intentional discrimination, carryover effects of prior discrimination, and disparate impact upon various demographic groups, including English-language learners. In addition to discriminatory claims, other due-process issues have included insufficient notice of the test requirements and lack of opportunity to learn the test content. These allegations have been put to the legal test in several states, including Alabama, California, Florida, Georgia, Illinois, and New York—with varying outcomes.

The requirement that a student pass an exit examination in order to receive a high school diploma is perhaps one of the highest stakes for which an exam can be used. The importance of a high school diploma practically guarantees that such a requirement will be tested in the courts. In fact, several states have preceded California in the implementation of high school exit examinations and their experiences can provide some insight as to the likely legal challenges that can be anticipated.

The decision in the *Debra P. v. Turlington* (1981) case in Florida established three standards that high school exit examinations must meet to stand up to constitutional scrutiny:

1. The test must measure knowledge and skills that are taught in the state's schools (referred to as "curricular validity")
2. Students must receive adequate notice of the test, the requirements for passing, and the consequences of not passing; an
3. The test must not intentionally discriminate against a protected group or class.

These areas also provide guidance for the bases of potential challenges, which include claims of inadequate content coverage, misalignment of curriculum to standards, insufficient notice, adverse racial/ethnic impact, and lack of accommodations for students with special needs and English-language learners.

Students Who Have Not Been Taught the Knowledge and Skills Measured By the Test

Clearly, examination content should be representative of what students have been taught. Alignment of curriculum standards and instruction to the examination is an essential first step in preparing for an exit exam. Curricular validity is considered to mean consistency between the test content and both what is found in the curricular materials and what is being taught in classrooms (NGA Reports Online, 1998). In *Debra P. v. Turlington* (1981), Florida demonstrated curricular validity through an outside study that surveyed curricular materials, teachers, district educators, and students and showed that the content on the exam was being taught in the classrooms.

California's *Guiding Principles for the High School Exit Examination Standards Panel* (California Department of Education, 1999a) describes the responsibility of school districts to prepare students to succeed. Alignment is the first item, indicating that California is aware of experiences in other states (e.g., Florida and Texas). The changes that must occur in the content and teaching methods require time. Thus, states need to provide for a sufficient timeframe between the introduction of the new test and the high stakes decisions to allow for the necessary alignment (suggested, for example, by Bond & King, 1995). This time period needs to be several years during which the districts and schools can work out and implement approaches that target fully preparing students for the examination.

The other prevalent concern is providing students with the opportunity to learn (OTL) what is tested. In *Goals 2000* (1994; 1996), "OTL standards" are defined as the "criteria for, and the basis of assessing the sufficiency or quality of the resources, practices, and conditions necessary at each level of the education system to provide all students with the opportunity to learn the material in voluntary national content standards or state content standards" (§3(a)(7)). The recommendation at the voluntary national level includes the following:

- Curricula, instructional materials, and technologies
- Teacher capability
- Continuous professional development
- Alignment of curriculum, instructional practices, and assessments with content standards
- Safety and security of the learning environment
- Non-discriminatory policies, curricula, and instructional practices
- Other factors that help students receive a fair opportunity to achieve the knowledge and skills in the content standards.

In efforts that have followed, the Joint Standards Panel (AERA, APA, & NCME, 1999) asserts that promotion and graduation tests should cover only the "content and skills that students have had an opportunity to learn" (146, Standard 13.5). *High Stakes* (National Research Council, 1999) recommends that "tests should be used for high stakes decisions about individual mastery only after implementing changes in teaching and curriculum that ensure that students have been taught the knowledge and skills on which they will be tested."

In addition to the overall implications of OTL, states must consider what this means for students with disabilities and English-language learners (ELL). The 1997 National Research Council study, *Educating One and All* edited by McDonnell, McLaughlin ,and Morison, and *High Stakes* edited by Heubert and Hauser (National Research Council, 1999), both state that if students with disabilities are required to pass state graduation tests, then educators must modify the students' IEPs so that the students get taught the relevant knowledge and skills. This is no small task, and a state needs to allow sufficient time to review and modify all its IEPs. For English-language learners, having been taught the relevant knowledge and skills takes on a different meaning when the test is in English. OTL for ELL means that these students have been given the chance to acquire not only the relevant subject matter but also the necessary levels of English proficiency. Again, a state must factor in adequate time for meeting these students' need. Other issues must be considered alongside alignment and OTL One of these is equal educational opportunity. Even if alignment of the content standards, instruction, and test content are in place and OTL actions were well established, a possible area of challenge to the examination is "adverse impact" based on claims of differential access to test preparation materials, substantial variations in setting and physical conditions during testing, and major disparities in quality of instruction (National Research Council, 1999). Questions about adverse impact should be included in validity studies, and the department of education, as the test user, should implement an evaluation component to keep track of the intended as well as unintended consequences of the examination (Messick, 1989).

Another important consideration related to the appropriateness of the content skills measured by the test is that it does not include items that are irrelevant to a construct or that underrepresent a construct. The issue is one of giving an unfair advantage or disadvantage to any particular subgroup. Adverse impact is a particularly relevant concern for subgroups such as minority students, students with disabilities, and English-language learners. These students will be at a disadvantage if the test measures facts or concepts that are not relevant to the content. As a guide to addressing these issues, Messick (1989) maintains that identification of either of the practices signals a validity problem, but absence of such identification indicates a policy problem. Thus, the test developer should provide evidence of the appropriateness of the content and cognitive processes to be measured. The department of education, as the test user, should include this in its evaluation component.

One of the biggest challenges for states is helping students who do not pass the graduation test, or who are in danger of not passing. Some states give funds to local school districts for remediation, which most often includes tutoring and extended learning time. Some states also fund summer school programs to assist students (NGA Reports Online, 1998). California has an initial statement related to this challenge in its *Guiding Principles for the High School Exit Examination Standards Panel* (California Department of Education, 1999a), where it posits that districts must provide alternatives for low-performing students, to include the reduction of electives to permit supplemental instruction during the academic year.

Insufficient Notice and Implementation

As a result of legal challenges in several states, the necessity to provide adequate advanced notice of the high stakes requirement of the test has become a fairly straightforward condition that all states now know must be an integral part of any graduation exam planning. In the case of the decision in *Debra P. v. Turlington* (1981), the court deemed four years as "sufficient notice." The basic language emphasizes that both students and school personnel should receive clear indications of the content and performance for which they will be held accountable. They also should be provided with general scoring guidelines and examples of proficiency in a standard. Other related materials (e.g., curricular frameworks, sample tasks, assessment specifications, and model responses) may be helpful in conveying expectations (National Research Council, 1999).

California has included such language in its law, requiring districts to provide early and adequate notice to students and parents about the test requirement. Further, the test becomes a requirement for graduation in the 2003–04 school year. The intent is that districts will be able to align curriculum and instruction to the state standards in all grades so students will be prepared to pass the exam prior to graduation ("SB 2X," 1999).

In addition to covering the advance notification requirement, the *High Stakes* report (National Research Council, 1999) stresses that states should take a number of steps before, during, and after test development to help ensure the appropriateness of the content and skills being measured. It recommends that the domain to be assessed on the graduation test should be clearly and widely publicized. These are important considerations that seem related to "notice," although they may not have been targets of direct challenges. The law is interested in whether tests are used inappropriately for student promotion, tracking, or graduation, and it is imperative that states be unambiguous about the purpose (National Research Council, 1999). States must fully and clearly define the function of an assessment and allow adequate time and resources to work out technical problems, give the test a fair trial, and gain support for it (Smith, Heinecke, & Noble, 1999c).

According to the policy researchers involved with the Arizona Student Assessment Program (ASAP), Arizona's experiences provide an interesting story, and perhaps guiding lessons, about the political and educational interplay on high stakes testing and the need to explicitly describe the function of an assessment (Smith, Heinecke, & Noble, 1999b). On the surface—that is, the official descriptions of the program and each of its elements at its inception in 1990—ASAP seemed clearly delineated. ASAP's original aims were grounded in educational reform, which involved using the state's existing standardized testing more as a diagnostic tool rather than as an accountability measure and altering instructional practices in classrooms. However, the legislative language did not mention ASAP, any particular form of testing (e.g., performance, authentic, alternative), or any principles of practice (Smith, Heinecke, & Noble, 1999d). According to the policy researchers, what happened is that the Legislature heard "accountability" (although this was minimized publicly), and the Department of Education heard "instructional reform." No funds were committed for teacher training in the reform principles and practices, so generally only school districts with an interest in this new direction and with sufficient wealth undertook staff development for ASAP. Consequently, the program moved away from a reform premise to a legislative one, and key legislators pressed the Department to start producing data for accountability

purposes. This set in motion a very constrained timeline that did not allow for building the capacity of schools and teachers, developing sound psychometric instruments, ironing out administrative wrinkles, conducting a study that compared ASAP with prior testing programs, or establishing an independent evaluation of the performance assessments (Smith, Heinecke, & Noble, 1999d). Thus, ASAP was viewed differently by various groups, who formed their perspectives only from direct experience. This conceptual confusion led to a range of viewpoints, including seeing as a set of performance tests, a measure of how students learn, a set of reform ideals, an authentic assessment, integrated learning, and a way out of the problems of high stakes standardized testing. Awareness of the accountability aspects of ASAP arose later (Smith, Heinecke, & Noble, 1999c), and the high stakes function was revealed fully when the state newspaper published the 1993 results by school and grade level and ranked them similarly to the way it had reported standardized test results (Smith, Heinecke, & Noble, 1999d). The ambiguity of the purpose of the ASAP and the multiple meanings led to the suspension of the test in 1995, just a week before it was scheduled to be administered (Smith, Heinecke, & Noble, 1999a).

Students of Color

Adverse impact is at the heart of challenges under civil rights laws, and was the basis for a case against the Texas Assessment of Academic Skills (TAAS) (*G.I. Forum and Image de Tejas v. Texas Education Agency, 1997*). A large gap exists in pass rates between white and minority students on the TAAS (Natriello & Pallas, 1999). Graduation rates for all students dropped precipitously after implementation of the TAAS. During the 8 years of its use, however, the White graduation rate has returned to its previous level, while graduation rates for Latino and African American students have not ("The case," 1999; Shannon, 1999). The judge recently ruled in favor of the defendants in this case, concluding that the state had demonstrated "educational necessity" ("Court rules," 2000; "Decision," 2000). Although the judge did not dispute plaintiffs' factual claims and acknowledged that the TAAS did have legally meaningful disparate impact against African American and Latino students, he concluded that the high stakes test approach is the only way to force improved outcomes in Texas education.

Although it is difficult to argue with the notion of school accountability for students' learning of basic skills for graduation, the test scores often take on such importance that negative practices may occur. In efforts to achieve higher scores, schools may retain students who are having difficulty in the grade level preceding the target test year. While the intended consequence is to provide the students with an added year to build their skills, the unintended consequence is that these students may be more inclined to drop out, which reduces the number of students in the test-taking pool and, consequently, the number who are likely to fail the test ("The case," 1999). Important considerations are setting the passing score based on accepted measurement procedures and using multiple measures to determine graduation (National Research Council, 1999; American Educational Research Association et al., 1985:54, Standard 8.12).

Natriello and Pallas (1999), who suggest that the growth of testing programs may be viewed from a social perspective as a control on awarding high school diplomas and on the related upward class movement of the disadvantaged, state, "the current testing boom may be

viewed both as an attempt to control the educational bureaucracies and to slow the rate at which new groups enter positions of power and influence in U.S. society." (p. 3) The essence of this problem is captured by Richard Jaeger, who testified in a North Carolina case that grade retention does not lead to improved learning but does increase the dropout rate ("N.C. lawsuit," 1997). Whether such a social view is acknowledged, there certainly is evidence that the recommendations of the Joint Standards Panel (AERA, APA, & NCME, 1999) are not always followed when tests are used in decisions about tracking (Darling-Hammond, 1991; Shepard, 1998), promotion (Shepard, 1991), and graduation (Stake, 1998). Such validity problems are a civil rights issue for minority students in particular, who are disproportionately assigned to lower track courses, retained in grade, and denied diplomas on the basis of test scores (Heubert, 2000). Over and over, these effects have been shown to have greater impact on minority students of color. Ultimately, the strongest practice to inoculate against such effects is to provide minority, and all, students with a high-quality education.

Actual charges of racial bias in graduation tests often focus on the method used for question or item selection. In the case of the TAAS, a type of correlation is used that typically selects questions (or items) with the highest correlations ("Racial bias," 2000). These are items with the greatest differences between high and low scorers. This method is frequently used to increase consistency and reliability. An unintended consequence is that this approach is likely to increase bias against minority students because, as a group, they do less well on the test overall.

When consistency is a primary factor used to screen test questions (or items), it is often the middle-to-upper-class experiences that are reflected, and certain other social and cultural perspectives are often eliminated ("Racial bias," 2000). One alternative is to select items via a different procedure such as an approach that reduces bias by using the smallest racial gaps from a particular subject area. This is more complex, however, when there are more than two racial groups to balance. Another possibility is to reject the premise of "unidimensionality" of this type of correlational selection. This assumes that test-takers use only one cognitive approach to solve a problem rather than multiple ways of thinking. If the basis of test development "assumes" unidimensionality but instruction "allows" for multi-cultural approaches there likely is cultural bias in the test ("Racial bias," 2000).

Score differentials between white and non-white students have come to be expected as a challenge in assessment programs. On most achievement and attainment indicators there are gaps. The usual, and most obvious, explanations point to inequalities in income and school resources. States are concerned about how to best address the difficulties without raising other charges leveled at graduation tests such as that they encourage "teaching to the test," which, in turn, is considered to have detrimental effects on disadvantaged and high minority schools and districts. An interesting direction in this area may be offered by Bempechat (1998), whose five-year study examined successful students to identify actions schools and families can take to help at-risk students reach their full potential. The factors behind the success of at-risk students are related to their beliefs about effort and ability, which form a fundamental foundation for academic success. Her study found that that "the overall pedagogical philosophy of Catholic schools is one that leads children to believe in their intellectual abilities and strive for academic excellence." This benefit was noted particularly

in increased academic success of African American and Latino students in Catholic schools. The key issue for educators is to identify strategies that will positively affect the way children think about success and failure (Bempechat, 1998).

Students with Disabilities

Federal and state laws now mandate inclusion of all students, including those with disabilities and English-language learners, in large-scale assessments(National Research Council, 1999). *Goals 2000*, *Educate America Act* (1994; 1996), Title I of IASA (1994), Individuals with Disabilities Act (IDEA, 1988, 1998), Section 504 of the Rehabilitation Act (1973), and Title II of the Americans with Disabilities Act (ADA, 1990) influence how students with disabilities are educated as well as included in assessment programs. They also establish legal rights for special needs students that can affect exit examinations.

Traditionally, students with disabilities have not participated in large-scale assessments. Parents' and educators' concerns have ranged from the stress of testing, confusion about modifications and accommodations, and the lowering of schools' average scores, to the possible mismatch between test content and curricula and difficulties in administering tests to students with severe disabilities (National Research Council, 1999). However, the importance of a high school diploma to one's future in our society has become increasingly clear and points to the need for more research to address the divided camps on the practice of issuing an alternative or modified diploma or certificate of completion to students with disabilities. Evidence is needed to help sort out charges of stigmatization to the students on the one hand and devaluation of the credential on the other (National Research Council, 1997).

Parents and educators of special needs students have indicated that they want special needs students to meet the same high standards established for general students (Ysseldyke et al., 1998). Given a greater push toward this goal, states will need to consider even more closely the "sufficient notice" component in their planning for graduation tests. As they struggle with testing systems that accommodate students, educators are recognizing that preparing for and establishing the changes to a student's individualized education program (IEP) that will provide the necessary support for inclusion in a state testing program, requires even more time than implementing other types of support programs. The Indiana Civil Liberties Union filed a suit in 1998 alleging that the state provided neither sufficient notice to students with disabilities of the requirement for graduation nor adequate accommodations in administering the tests to special needs students (Solida, 2000).

Nevertheless, with current graduation exams, as with most state assessment programs, students receiving special education services frequently may be exempted from the test. For example, in Texas, the student's special education committee determines if the TAAS or end-of-course examinations are appropriate measures of the student's progress. Students who are not exempt based on their special education status (e.g., learning disabilities such as dyslexia) may be eligible for testing accommodations (Natriello & Pallas, 1999). In New York, students with disabilities also may be exempted from the Regents Examinations, or may be eligible for testing modifications. In Minnesota, students with disabilities are eligible for accommodations or modifications when such specification is included in the individualized

education program (IEP) or 504 plans. Modifications that involve adjustment in the test that actually changes the standard result in a notation on the transcript that indicates "Pass-Individual" rather than "Pass-State" (Natriello & Pallas, 1999).

Among the 12 recommendations made in the study group report to the National Research Council, *Educating One and All*, (National Research Council, 1997; pp. 197–210), the following seem particularly worthy of consideration in California:

- States and localities that decide to implement standards-based reform should design their common content standards, performance standards, and assessments to maximize participation of students with disabilities.
- The presumption should be that students with disabilities will participate, but an individual student may require alterations; such decisions must have a compelling educational justification.
- States and localities should revise policies that discourage maximum participation of students with disabilities and provide incentives to encourage widespread participation.
- When content and performance standards are altered for a student with a disability, (a) the alternate standards should be challenging yet achievable, (b) they should reflect the full range of knowledge and skills that the student needs to live a full, productive life, and (c) the school system should inform parents and the student of any consequences of these alterations.
- Accommodations should be justified on an individual basis and should be unrelated to the knowledge and skills being measured.
- States and localities should provide information to parents to help them make more informed choices about their children's participation and to understand the consequences of the choices.
- Students with disabilities should be given sufficient opportunity to learn the knowledge and skills expected of them before high stakes consequences are attached to their performance.
- Policymakers should monitor the unintended consequences of participation in standards-based reform for all students including those with disabilities.
- States should design standards policies that reflect realistic timelines and the resource levels needed to implement standards-based reform.

The validity of an exit exam that allows for accommodations for students with disabilities is a major concern, but one for which there is little guidance from research. Accommodations are meant to correct for score influences that are caused by a disability unrelated to the construct being measured. However, the threat is that the correction may be the wrong one, irrelevant, or excessive. Many approaches to assessment accommodations do assume that disabilities are not directly related to the construct being tested, and such relationships are very important to judging the validity of inferences based on the scores. Inconsistent descriptions and classifications of disabilities and lack of clearly segregated skills have resulted in difficulties in maximizing validity when accommodations are involved (National Research Council, 1997). More research on the validity of scores from

accommodated testing is needed. The 1995 field test of the National Assessment of Educational Progress (NAEP) in mathematics and science provided some evidence about the relationship of accommodations to standards-based reform. The results revealed that it is feasible to assess a greater number of students with disabilities than previously. However, given small sample sizes and lack of clarity and consistency between disability and corresponding accommodations, the results are only suggestive of the effects of accommodations on score comparability (National Research Council, 1997).

English-Language Learners

Failure to graduate from high school has tremendous consequences on a student. Hauser (1997) reports evidence that such failure is linked to problems of employment, earnings, starting and maintaining a family, civic participation, and health. The Mexican American Legal Defense and Educational Fund (MALDEF) pointed to such consequences when it filed a lawsuit over the Texas graduation exam, alleging that it discriminates against Hispanic and African American students, who fail the test in disproportionate numbers (Smith, 1999; Herrick & Walt, 1997). Whether failure to gain a high school diploma is due to graduation tests or other reasons, Hauser (1997) aptly captures the bleak outlook for these students: "Failure to obtain at least a high school diploma looks more and more like the contemporary equivalent of functional illiteracy. High school dropout indicates a failure to pass minimum thresholds of economic, social, or political motivation, access, and competence" (p. 154).

In the 1998–99 school year, approximately 37% of California students were not native-English speakers; almost 25% were classified as ELL (English-language learners) (California Department of Education, 1999b). Thus, the preparation and performance of these students on the graduation exam are very important considerations for California.

Several current federal laws stipulate the participation of *all* students in assessments of performance (e.g., Goals 2000, Title I (Helping Disadvantaged Children Meet High Standards), and Title VII (Bilingual Education)) (National Research Council, 1999). The primary difficulty with assessment of English-language learners (ELL) is obtaining accurate reflections of their knowledge in a given subject, because deficits in language may impede the students' understanding of—and response to an exam question. Therefore an incorrect response may not reflect a lack of content knowledge, but rather a language problem. Consequently, the first step of the "decision" is whether the ELL student should participate in the assessment at all—that is, whether the student should be exempt. Such decisions are usually based on some indicator of English proficiency.

California decided to allow fewer exemptions for the SAT9, which is required of all its students grades 2 through 11. It requires all ELL who have been in the state for more than 1 year to take the test in English (National Research Council, 1999). However, the superintendent of San Francisco Public Schools filed a suit against the policy, and prevailed, maintaining that it takes more time than has been allowed before a student becomes sufficiently proficient to take the test in English ("S.F. files suit," 1998).

For the high school exit examination, California will require all students who do not have waivers to pass in order to receive a diploma. ELL, as a designation, is not a reason for a waiver; thus, it would be more accurate to use a term such as "temporarily exempt" or

"deferred" for ELL (as clarified during meeting of California Board of Education, March 9, 2000). Nevertheless, an essential consideration for testing ELL is whether the students have been taught the knowledge and skills that the test measures, including oral and written English.

When an ELL student is included, the state must address the best method of administering the test or modifying testing procedures to obtain accurate measures. Florida, Texas, and New York, which also have large ELL populations, have addressed some of the challenges in the following ways (National Research Council, 1999).

Florida

- ELL in approved program fewer than 2 years may be temporarily exempted from taking the high school competency test, but they cannot receive a standard diploma until the test is passed.
- Students not receiving special language assistance may not be exempted solely on basis of ELL designation.
- ELL who do not pass the competency test can return for a 13th year to focus on sections that were not passed; those who do not pass can opt for a certificate of completion rather than a diploma.

New York

- Students must pass state Regents exams or approved alternative assessments in four core courses—English, mathematics, social studies, and science—plus other courses, to demonstrate achievement of state standards.
- ELL who enter U.S. schools in 9th grade or later can take all required Regents exams, except English/Language Arts, in five different languages.
- All ELL must pass the English Regents exam to receive a diploma.
- State is implementing a comprehensive strategy to help ELL pass the exam, which includes (a) identification of model programs; (b) professional development for all teachers of ELL on integration of English language arts standards into instruction; (c) alignment of ELL programs with state standards; and (d) other added instructional programs and resources, publications, and statewide symposia.

Texas

- Students must pass an exit exam in reading, writing, and mathematics, taken in 10th grade, to graduate.
- School-based committee uses six criteria to determine whether each ELL is tested on the English TAAS or Spanish TAAS, or is exempted and given an alternative assessment.
- ELL entering U.S. schools in third grade or later are required to take English TAAS after 3 years.

Aside from commonly expected validity problems of bias in testing (e.g., small number of ELL in sample, content bias, and linguistic and cultural biases), when a test is developed in a non-English version, other concerns arise, including translation and score equivalence.

A basic truism is that to some extent all assessments are measures of language (see for example, National Research Council and Institute of Medicine, 1997: 120–122 and The National Education Goals Panel, 1998: 6). However, in spite of potential concerns and problems with native-language testing, given California's demographics, this is an important consideration for its graduation exam. Lack of proficiency in the language of a test can severely underestimate the test taker's knowledge. Further, scorers may introduce errors on extended response or performance-based items through unwarranted influence of linguistic elements of responses (National Research Council, 1999).

There are clear calls for more research that would inform decisions about obtaining valid scores for ELL in large-scale assessments (National Research Council and Institute of Medicine, 1997; Olson & Goldstein, 1997). Such research would help address existing difficulties in various strategies being used or considered to enhance ELL participation in large-scale assessments (National Research Council, 1999). For example, use of native-language assessments requires more research on equivalency, validity and reliability before a state incorporates them in its assessment program. Information from New York, which is trying native-language exams in all subjects except English, will help inform this area.

A pressing need in the area of English-language learning and assessment is for a definition of English proficiency that can be applied across the country (Olson & Goldstein, 1997). Issues related to this include the following:

- Use of an English-language assessment of proficiency rather than years in English-only instruction; years of instruction may not accurately predict literacy.
- A proficiency measure should include both oral and written language.
- An individualized approach to match the testing to a student's English literacy level requires development, validation, and adoption of a standard procedure to determine levels of literacy; then, a threshold level would be established to indicate when a student could take the standard English assessment.
- For subjects other than English, when a student is proficient in more than one language, and tests are available in those languages, the student's more/most proficient language should be the basis for testing.
- The amount of instruction in a particular subject and in the native language should be considered in deciding the language of the assessment. (Texas' procedure for this type of decision-making will help inform this area.)

Currently, California statute (Senate Bill 2X) includes language that directs the school district to make a determination of English-language proficiency and whether a given student should be assessed by the graduation exam or deferred for up to 24 months to allow the student at least 6 months of English-language instruction. Under the law, a district must offer summer school programs to any student who does not demonstrate sufficient progress toward passing the graduation exam (using STAR results or grades and other academic indicators). Funding for this initiative includes ELL who need to build their English-language proficiency ("SB 2X," 1999). This is an important step; however, based on experiences in Texas, legal challenges may invoke inequities in the entire school system for

students of color and ELL (Herrick & Walt, 1997). This suggests the need for a state to examine its educational infrastructure and to begin a standard of instruction from the beginning of all students' education.

The Legal Milieu

These groups of students—students who have not been taught the knowledge and skills measured by the test, those who claim insufficient notice, students of color, students with disabilities, and English-language learners—are all groups who may eventually feel that they have a legitimate legal challenge to the California High School Exit Exam. Certainly proactive steps can be taken to minimize the likely challenges. Unfortunately, there have been inconsistent and even contradictory legal decisions in the past, and there is no single set of rules or guidelines that can guarantee successful legal defense.

In the face of probable legal challenges and decisions, well-informed test implementers would do well to design and build a test framework with these issues in mind. The educational community has vast experience with considering issues of adverse impact and the needs of various student groups. Psychometric, legal, and education expertise all play critical roles.

Summary

As California undertakes the development and implementation of its high school exit examination, policymakers and assessment staff are interested in using the experiences of states that preceded them in this endeavor. This background chapter focused on the most common issues and concerns and the current "state-of-the-art" in high stakes graduation testing. There are two important paths that a state must traverse in building its exit exam program. It is clear that a state must consider critical links, or logical steps, among the parts of its program in wisely planning and establishing the foundation and structure that will support its graduation test. A state can also benefit by paying attention to legal decisions in other states, which have been a predominant influence in setting the requirements a state must meet in creating its exit examination.

Critical Links

A strong system of supporting and integrated "links" is necessary to achieve the desired outcomes of graduation testing. The most common reasons for establishing an exit exam program are to promote standards for awarding a diploma and improved classroom instruction. However, these improvements do not grow from the decision itself; they require planning and nurturing. The following test criteria are critical links for the state to consider:

- *The Standards of Test Use…*
  - Have measurement validity, defensible test psychometrics, and appropriate content aligned with standards and taught throughout the state;
  - Address causes of poor results with valid explanations; and
  - Lead to educationally beneficial outcomes.
- *The Test Scores…*

  - Consider use of test scores for diagnostic purposes, or
  - Couple "pass-fail" scores with program of diagnostic information and appropriate remediation.
- *The Score Reports…*
  - Take into account individual- and aggregate-level test results;
  - Reflect audience needs and uses clear, straightforward language; and
  - Provide a well-designed format.
- *Public Awareness and Communication by the state…*
  - Attend to the dissemination of explanatory and promotional information about the examination and its related components; and
  - Use a variety of dissemination and distribution formats and strategies appropriate to the audience.
- *The Instructional Changes…*
  - Involve relating patterns of test outcomes to appropriate instructional approaches;
  - Take into account individual- and aggregate-level results;
  - Address instructional needs of students who fail the exam;
  - Deliver appropriate and timely first-time instruction to students in advance of taking the exam;
  - Establish sound feedback mechanism between results and educators; and
  - Provide adequate resources for materials and teacher training to facilitate desired instructional requirements and modifications.

Legal Influences

Policy decisions for any educational area are first and foremost established and manipulated through laws and statutes. Every day, education policymakers deal with and interpret federal, state, and local mandates in regard to numerous aspects of educating children. These regulations act as a foundation and scaffolding that must be solidly in place to support implementation of new policies and programs, such as graduation testing. For example, Titles I and IX have long functioned to specify education requirements, including those that apply to testing in general. Civil rights regulations also must be considered in educational decision-making because students in identified groups retain their rights in all aspects of their lives, including education. Policymakers must ensure adherence to these laws and safeguards in all educational areas. Thus, any of these laws can serve as the basis for legal challenges in education.

In the area of testing, the issues focus on whether the test may be discriminatory or used inappropriately for promotion, tracking, or graduation. Unfortunately, there is no absolute judgment on what constitutes discriminatory or inappropriate practices. This results in decisions on analogous legal challenges that may be diametrically opposed, or that finish in similar conclusions but are based on different justifications. However, the Improving America's Schools Act (IASA) of 1994, which made major changes to Title I, is nearing its

2000–2001 school year date for implementation of the new testing system. States must interpret and operationally define the requirements of the act, and it seems likely tests used in this area will become a target of legal challenges.

Another side of legal considerations is how courts have applied legal principles to issues related directly to students and high school graduation. Any state contemplating implementation of an exit examination is keenly aware that legal challenges will most likely follow. It makes sense, then, to study what has happened in other states in order to prepare for or try to avoid legal action. Legal decisions in other states have served as one primary source of establishing basic requirements of graduation tests. Based on the decision in Florida, there are three standards that high school exit examinations must meet to pass constitutional muster:

1. The test must have curricular validity, that is, measure knowledge and skills that are taught in the state's schools.
2. Students must receive adequate notice of the test, the requirements for passing, and the consequences of not passing.
3. The test must not intentionally discriminate against a protected group or class.

Although these represent results in a particular state, they are currently the accepted "starting point" for other states in their planning, development, and implementation of an exit examination. Texas has experienced the most recent decision related to its graduation test, and the state education agency prevailed on the basis of supporting each of the areas above. Therefore, it would seem that the most prudent recommendations to California and other states regarding graduation testing are to (a) adhere to all the applicable laws related to education, students, and programs generally, and to testing specifically; (b) follow the related professional standards for developing, implementing, monitoring, and use of the test; and (c) document the processes and results related to each aspect of the laws and the test.

Perhaps the most useful guidance for California and other states comes from U.S. Secretary of Education, Richard W. Riley, a strong supporter of standards-based reform. In his February 22, 2000 "State of American Education" address, Riley called for a "midcourse review" of the standards movement, a step he said was needed "because there is a gap between what we know we should be doing and what we are doing" (Riley, 2000: 6). Specifically, Secretary Riley said that state standards should be "challenging but realistic…. [Y]ou have to help students and teachers prepare for these [high stakes] tests—they need the preparation time and resources to succeed, and the test must be on matters that they have been taught" (Riley, 2000: 7). He also advised states not to rely on any single measure of students' knowledge in making high stakes decisions: "All states should incorporate multiple ways of measuring learning" (Riley, 2000, 6).